

Fig 1c: Pass Rates with 95% Confidence Intervals

Agent	Model	Pass Rate (%)	95% CI	N Trials
Terminus	gpt-5-mini	48.76	[44.04, 53.71]	445
Terminus	gpt-oss-120b	52.36	[47.63, 56.85]	445
Terminus	gpt-oss-20b	25.39	[21.35, 29.21]	445
Terminus	gpt-5-codex	63.15	[58.65, 67.64]	445
Terminus	gpt-5-nano	33.48	[29.21, 37.75]	445
Terminus	gpt-5	65.84	[61.34, 69.67]	445
Codex	gpt-5-mini	38.65	[34.16, 42.93]	445
Codex	gpt-oss-120b	44.94	[40.00, 49.66]	445
Codex	gpt-oss-20b	20.67	[16.85, 24.27]	445
Codex	gpt-5-codex	56.40	[51.91, 60.68]	445
Codex	gpt-5-nano	26.52	[22.25, 30.56]	445
Codex	gpt-5	57.98	[53.03, 62.25]	445